Feature wise representation for both still and motion 3D medical images

Xiaohong Gao
School of Science and Technology,
Middlesex University, London, NW4 4BT, United Kingdom
x.gao@mdx.ac.uk

Abstract— With the arrival of the state of the art medical imaging equipment, a plethora of images are acquired not only in higher dimensions (3D+) but also with various presenting forms of either still or motion, complicating data management systems even further. This paper offers, from an application point of view, representations of content features from both still 3D MR brain images and 3D ultrasound cardiac video sequences by demonstrating a developed online content-based image retrieval system, MIRAGE. The approaches of 3D SIFT coupled with sparse code have been appointed to facilitate the representation of image features, whereas widely applied four texture based approaches are also implemented to allow users' benefit of different retrieving intentions.

Keywords-Imaging, representing biomedical knowledge, multi-media image retrieval.

I. INTRODUCTION

Three dimensional images, including both still pictures and motion videos, are at present a common form in hospitals in assisting clinicians performing diagnosis or intervention, such as 3D MR brain images or ultrasonic 2D video clips with time scale being the third dimension. With regard to still images, structured templates usually exist, containing constructive geometric properties, such as shape or texture, the characteristics that can be employed to standardise those images in terms of their content features. For motion videos that are of a function of both space and time, however, additional data might be needed to ensure that the starting time is in the same cycle for all the datasets in a database.

With regard to 3D still brain images, many methods for feature representation of content have been developed, including intensity-based [1], physiological-information-based [2], and textured-based [3]. With the application of the Talairach [4] brain atlas, intensity-based approach takes advantages of spatial references to concert a region-based retrieval, in which a regional or a volumetric data is expressed as <x, y, z, value>. Since this representation is reminiscent of image matrix, depending on the size of each region/volume, the presentation volume of each image can be as large as the size of an image dataset itself, to a certain extent, missing the point of representing images using features to reduce redundant information and therein being prone to artefacts. On the other hand, by using physiological-information-based retrieval, a number of semantic contents can be dealt with by employing physiological kinetic features of images [2]. Although effective, this method is very discipline-constrained and heavily relies on the additional supply of extra information, such as blood samples in order to derive plasma time activity (PTA) curves, gaining prospective in searching for alternative approaches.

Texture-based approach to extract features from 3D MR brain images is originally applied by the employment of 2D Gabor filter [5], which in essence remains a 2D approach. After further extension and tailoring, 3D textures are obtained from four well known approaches in [3], including Local Binary Pattern (LBP), Grey Level Co-occurrence Matrices (GLCM), Wavelet Transforms (WT) and Gabor Transforms (GT). For the application to 3D brain images using texture-based approaches, the prerequisite resides on the spatial normalisation to a standard template. In this way, the comparison of each sub-volume (cube) within the same location between images can be conducted under the assumption that geometrically, all brains bear similar anatomic structure. In practice, however, due to lesioned brains sustaining appreciable distortions, the texture features even from normalized images cannot be unique in relation to representations, which is also true for video images. To overcome this deficit, in this study, key point based feature representation by the appointment of scale invariant feature transformation (SIFT) coupled with a machine learning technique of sparse coding is implemented, towards developing an online system of content-based image retrieval (CBIR) system for medical images (http://image.mdx.ac.uk/time/demo.php).

SIFT descriptors [6], being invariant to geometric transformations of translation, scaling and rotation, provide robust feature matching mechanisms across a substantial range of changes in illumination while with the presence of noise. It is therefore widely applied in the domain of object recognition and image stitching. In addition, a 3D extension of the SIFT algorithm has recently been proposed in the literature on 3D volumetric data analysis, such as, action recognition in video volumes [7], object recognition in CT complex volume [8], 3D medical registration and panoramic medical image stitching [9]. Evidently, each type of dataset present their own unique set of characteristics that subsequently require the application of SIFT in a different way as to the extraction of potential features.

The remainder of the paper is structured as follows. Section II addresses mathematical formulae underling the methodology, which is then followed by the results on the application of CBIR for brain images and classification for cardio videos given in Section III. The paper is then concluded in Section IV.

II. METHDOLOGY

The procedures involve sparse coding of 3D salient content features and the creation of a codebook of visual dictionaries.

A. Pre-processing

Local visual feature selection usually remains the first task to conform. For an MR brain image, after spatially normalised to an MNI (Montreal Neurological Institute) template, it is spatially partitioned equally into non-overlapped sub-cubes of 512 (=8×8×8). In this way, the retrieval practice can be conducted by focusing on each corresponding sub-cube without the need of cross-comparison, saving considerable retrieving time.

On the other hand, for motion video clips, low-level interest-points are found out first as potential candidates for cross-image comparisons. To detect a spatial-temporal interest point, the technique of Cuboid detector [10] are opted for in an attempt to overcome the inability that many other methods suffer from with reference to incorporating temporal information.

With respect to still brain images, 3D SIFT descriptors are applied to enumerate local visual features by computing a 3D gradient orientation histogram for each of 512 sub-volumes. In doing so, further division is performed on each sub-volume to create eight (=2 × 2 × 2) sub-blocks, upon each of which, the magnitude and orientation of its gradient are calculated, by the application of 1D Haar wavelet transform in each of x, y and z directions respectively, forming bins of orientations accumulating the magnitude values that share the same gradient orientation. Subseuently, based on the tessellation technique, eighty bins of orientation are rendered in a 3D orientation sphere for each sub-block, leading to a feature vector of 640 dimensions (=2×2×2×80) for each sub-volume.

Likewise, a 3D SIFT descriptor consisted of 640 elements is extracted as a feature vector in video images, in this case to represent each interest voxel instead of sub-volume. Centered at each interest voxel, detected by using the Cuboid detector, a 12 x 12 x 12 volume of neighbourhood is selected and then divided into 8 (=2x 2 x2) sub-volumes that subsequently undertake the same procedure of calculation of gradients as explained above.

B. Visual Vocabulary Construction Using Sparse Coding

Once 3D SIFT features, i.e., the candidates for unit elements, or "words" in a visual dictionary, are accounted for from each sub-volume or voxel, sparse coding follows to allow the creation of dictionary of visual 'words'.

In theory, by modeling data vectors as a sparse linear combination of a set of basic elements or 'words', sparse coding [11] encodes each descriptor of an image by solving the optimization problem as formulated in Eq.(1).

$$\min_{U,V} \sum_{m=1}^{M} \|x_m - Vu_m\|_2^2 + \lambda \|u_m\|_1$$
$$\text{Subject to}: \|v_i\| \le 1, \quad \forall i = 1,\ldots,K \tag{1}$$

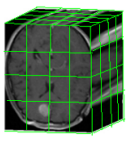
where $X = [x_1, x_2, \ldots x_M] \left(x_m \in R^{dx1}\right)$ refers to a set of 3D SIFT descriptors as described above from a 3D training dataset. $V = [v_1, v_2, \ldots v_K] \left(v_i \in R^{dx1}\right)$ indicates the K bases, also known as the dictionary or codebook; and $U = [u_1, u_2, \ldots u_M] \left(u_m \in R^{Kx1}\right)$ denotes sparse codes for images based on codebook V. In addition, M refers to the number of total training samples in the training dataset, whereas λ indicates the constant coefficient that is generated by L_1 norm ($\|\cdot\|_1$) regularization, which acts as the penalty function to produce sparse coefficients and being robust to irrelevant features.

C. Max-pooling for Image Representation

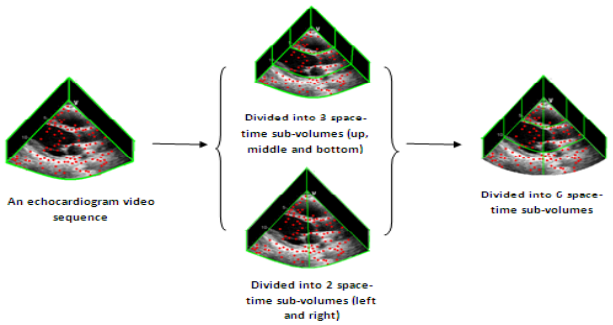
To take into account of spatial location of local features, a pooling technique is applied to the SIFT sparse code that is calculated from different regions of an image to create the representation of an image by concatenation. In this study, max-pooling technique is employed as opposed to average-pooling by choosing the max value from a set of inputs as illustrated in Eq. (2).

$$z_i = \max\left\{|u_{ij}|, j = 1,2,\ldots,S\right\} \tag{2}$$

where u_{ij} in Eq.(2) indicates the i^{th} ($i \in [1, K]$) 3D SIFT sparse code for the j^{th} ($j \in [1, S]$) sub-volume or interest point within a pooling region, with K indicating the size of codebook V, whereas S refers to the total number of the sub-volumes in each pooling region. Figure 1 schematically demonstrates the pooling process, i.e., an MR cubic image being equally divided into 64 (= 4 × 4 × 4) regions, whilst 12 pooling regions are drawn up from a clip of ultrasonic videos to acknowledge the differences of the structure of a moving heart at three levels.



(a) Pooling regions for 3D brain image



(*b)Pooling regions for echocardiogram video*

Figure 1. The pooling regions for 3D MR brain images (a) and echocardiograms (b)

D. Comparisons between Datasets

For still images, both histogram intersection and Chi-square histogram are applied to measure the degree of similarity between two images. Whereas for video images, the comparison is based on viewport classification which is performed using an approach of multiclass SVM (Support Vector Machine) with a linear kernel that is calculated in Eq. (3).

$$k\left(\overline{F}_i, \overline{F}_j\right) = \overline{F}_i^T \overline{F}_j \tag{3}$$

where $\overline{F}_j$ is the feature representation of video *j*. With regard to binary classification, an SVM aims to learn a decision function based on the training dataset, which is defined in Eq. (4).

$$f\left(\overline{F}\right) = \sum_{i=1}^{n} a_i k\left(\overline{F}_i, \overline{F}\right) + b \tag{4}$$

In order to obtain an extension to a multi-class SVM, the trained videos are represented as $\left\{\left(\overline{F}_i, l_i\right)\right\}_{i=1}^{n}$, where $l_i \in \{1,2...L\}$ denotes the class label of trained video *i*. One-against-all strategy is applied to train the total number of L, the binary classifiers.

III. RESULTS

A. Dataset Collection

In this investigation, the data at our disposal are of both 3D MR still brain images and 3D (i.e., 2D video) echocardiographs, all in the format of DICOM. The characteristics of these datasets are given in Table 1.

TABLE 1. THE INFORMATION RELATED TO THE DATASETS AT THE DISPOSAL.

	Still Brain Images	**Ultrasonic Motion Video Cardiac Clips**
Normal	34	14
Abnormal	86	58
Subject	120	72
Resolution	500×500×45 (mm^3)	434 × 636 ($pixel^2$) × 26 (frames)
Imaging Tool	GE Genesis_Signa 15000T (1.5-T) whole-body MR	GE Vivid 7 Ultrasound
Total	120	219

Additionally, the ground truth is based on the locations of lesions from brain images and eight viewpoints of heart video clips, which are marked by clinicians.

B. Training

In the training stage, a 3D brain codebook composed of 64 sub-codebooks is obtained from 500 descriptors randomly selected from 960 sub-volumes.

With regard to echo-cardiac video images, 80,000 interest points firstly detected by Cuboid detector are randomly selected from the video clips (n=219) as a training dataset for the generation of the codebook.

C. Comparison Results

As discussed in the INTRODUCTION, 3D still images can also be represented using 3D texture-based approaches. Therefore comparison with the four popular texture approaches [3], is also carried out, with the comparison results given in Table 2, by which retrieval task focuses on the location of lesions by using the measure of mean average precision (MAP).

TABLE 2. MAP VALUES FOR THE FIVE 3D APPROACHES INCLUDING GLCM, WT, GT, LBP AND SIFT

Methods	**Without sparse coding**	**With sparse coding**	
		Histogram intersection	**Chi-squared histogram**
3D GLCM	0.3034	0.3291	0.3510
3D WT	0.3096	0.3375	0.3687
3D GT	0.3074	0.3863	0.3954
3D LBP	0.3308	0.4027	0.4012
3D SIFT	0.3959	0.4013	0.4098

As presented in Table 2, the approach of 3D SIFT that has been furthered in this research outperforms the other four with the average MAP value of 0.4098 according to Chi-squared histogram distance. In addition, the implementation of sparse coding improves the performance of all five approaches, specifically for 3D GT with MAP value increasing from 0.3074 to 0.3954, implying the significance of contribution of machine learning technique to the presentation of visual features.

For video images, the retrieval is based on the viewpoint. Therefore in essence the procedure is of a classification. Table 3 illustrates the confusion matrix of the classification based on eight standard viewpoints.

TABLE 3. CONFUSION MATRIX FOR 8 ECHOCARDIOGRAM VIEW CLASSIFICATION WHERE AR=ACCURACY RATE, ER=ERROR RATE, AND THE VERTICAL AXIS REFERS TO GROUND TRUTH

	Classification Results	**AR**

	A2C	A3C	A4C	A5C	PLA	PSAA	PSAP	PSAM	
A2C	**32**	2	6	0	0	2	0	0	0.76
A3C	6	**17**	6	0	0	3	0	0	0.53
A4C	5	1	**26**	0	2	0	0	0	0.76
A5C	1	0	2	**4**	0	0	0	0	0.57
PLA	1	0	0	0	**34**	2	0	0	0.92
PSAA	2	0	0	0	4	**28**	1	4	0.72
PSAP	0	0	0	0	2	5	**12**	0	0.63
PSAM	0	0	0	0	1	3	1	**4**	0.44
ER	0.32	0.15	0.35	0	0.21	0.3	0.14	0.5	

In summary, the average AR (AAR) for all classes is 72% (=157/219), and the average ER (AER) is 28% (=62/219). According to the data in Table 3, the most erroneous classification takes place within the classes having the similar view points, such as views taken from Apical angles (4 views) and Parasternal Short Axis (3 views). The unique view of PLA gives the best performance (AR=92%).

D. Content-based Image Retrieval System of MRIAGE

Figure 3 demonstrates a screenshot of the developed online retrieval system, MIRAGE at http://image.mdx.ac.uk/time/demo.php. Built upon the open source GNU Image Finding Tool (GIFT), the online system fosters the Query-by-Example (QBE) paradigm coupled with user-relevance feedback facility whereby retrieved images most closely resemble a query image in appearance. This system has been serving as an e-leaning tool for post-graduate students and has currently accommodated over 100,000 medical images of both 2D and 3D.

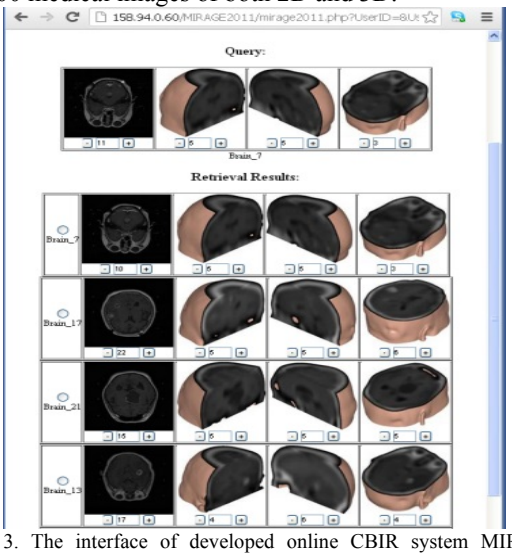


Figure 3. The interface of developed online CBIR system MIRAGE, showing the retrieval results of 3D brain images.

IV. CONCLUSION

This paper presents the promises of feature wise representation using the synergy of 3D SIFT and sparse coding, in an attempt to lead to the application of CBIR to medical images in a clinical sector. In particular, both still and motion images of different modalities are under investigation and are being implemented in the online CBIR system, MIRAGE, for the intention of educational at present. Reflecting from the feedbacks given by the students who have used the system, searching and learning medical images by way of CBIR has shown significant advantages, especially for those who are new to each image modality as well as medical anatomy. It is anticipated that similar work will be conducted on 4D echocardiograms in the new future. In addition, the missing links between the features and their semantics at higher level will be bridged in the future.

ACKNOWLEDGEMENT

This work constitutes part of the WIDTH (2011-2014) project that is sponsored by EC under Marie Curies Scheme. Their financial support is gratefully acknowledged.

REFERENCES

[1] M. Arya, W. Cody, C. Faloutsos, J. Richardson, and J. Toya, 1994. **"**QBISM: Extending a DBMS to Support 3D Medical Images**,"** Proceedings of the Tenth International Conference on Data Engineering, pp. 314-325, 1994.

[2] W. Cai, and D. Feng, "Content-based Retrieval of dynamic PET functional images," IEEE Transactions on Information Technology in Biomedicine . 4, 2 (2000), pp. 152-158.

[3] X. Gao, Y. Qian, M. Loomes, R. Comley, B. Barn, A. Chapman, J. Rix, R. Hui, Z. Tian, "Retrieval of 3D Medical Images via Their Texture Features," International Journal On Advances in Software, 2011, no 3&4, pp. 499-509.

[4] J. Talairach, and P. Tournoux, " Co-planar Stereotactic Atlas of the Human Brain," Thieme, Stuttgart, 1988.

[5] T. Glatard., J. Montagnat., I.E. Mgnn, "Texture based medical image indexing and retrieval: application to cardiac imaging," Proceedings of the 6th ACM Sigmm International Workshop on Multimedia Information Retrieval, 2004.

[6] D.G. Lowe., D.G. "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision (IJCV), 60 , 2 (2004), pp. 91-110.

[7] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition," IACM Conference on Multimedia, 2007, pp. 357-360.

[8] G. Flitton, T. Breckon, N. Megherbi, "Object recognition using 3D SIFT in complex CT volumes," British Machine Vision Conference (BMVC), 2010, pp. 1-12.

[9] D. Ni, Y. Qi, X. Yang, Y. Chui, T. Wong, S. Ho, and P. Heng, "Volumetric ultrasound panorama based on 3D SIFT," International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'08), 2008, pp. 52-60.

[10] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-temporal Features," IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), 2005, 65–72.

[11] H. Lee, A. Battle, R. Raina, A.Y. Ng , "Efficient sparse coding algorithms," Advances in Neural Information Processing Systems (NIPS), 2006, pp. 801-808.